Analysis of Airline Delay and Cancellation Data, 2009 - 2018

Paul Smith

Western Governors University

**Table of Contents**

[A. Project Overview 3](#_Toc114439248)

[A1. Research Question or Organizational Need 3](#_Toc114439249)

[A2. Context and Background 3](#_Toc114439250)

[A3. Summary of Published Works 3](#_Toc114439251)

[A3a. Relation of Published Works to Project 3](#_Toc114439252)

[A4. Summary of Data Analytics Solution 3](#_Toc114439253)

[A5. Benefit to Organization and Decision-Making Process 3](#_Toc114439254)

[B. Data Analytics Plan 4](#_Toc114439255)

[B1. Goals, Objectives, and Deliverables 4](#_Toc114439256)

[B2. Scope of Project 4](#_Toc114439257)

[B3. Standard Methodology 4](#_Toc114439258)

[B4. Timeline and Milestones 4](#_Toc114439259)

[B5. Resources and Costs 4](#_Toc114439260)

[B6. Criteria for Success 4](#_Toc114439261)

[C. Design of Data Analytics Solution 6](#_Toc114439262)

[C1. Hypothesis 6](#_Toc114439263)

[C2. Analytical Method 6](#_Toc114439264)

[C2a. Justification of Analytical Method 6](#_Toc114439265)

[C3. Tools and Environments of Solution 6](#_Toc114439266)

[C4. Methods and Metrics to Evaluate Statistical Significance 6](#_Toc114439267)

[C4a. Justification of Methods and Metrics 6](#_Toc114439268)

[C5. Practical Significance 7](#_Toc114439269)

[C6. Visual Communication 7](#_Toc114439270)

[D. Description of Datasets 8](#_Toc114439271)

[D1. Source of Data 8](#_Toc114439272)

[D2. Appropriateness of Dataset 8](#_Toc114439273)

[D3. Data Collection Methods 8](#_Toc114439274)

[D4. Data Quality 8](#_Toc114439275)

[D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance 8](#_Toc114439276)

[D5a. Precautions 8](#_Toc114439277)

[E. Sources 9](#_Toc114439278)

# Project Overview

# A1. Research Question or Organizational Need

This project will analyze airline delay and cancellation data from 2009 to 2018 and determine the causes of flight delays. Flight delays cost the US economy a great deal of money ($32.9 billion in 2007) and determining the cause of these delays could allow this cost to be mitigated or reduced.

# A2. Context and Background

A 2010 report sponsored by the Federal Aviation Administration (FAA) analyzed a variety of cost components caused by flight delays. This included the cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The report concluded that the total cost of all US air transportation delays in 2007 was $32.9 billion. Clearly, flight delays are a serious and widespread problem in the US.

An exploration of multi-year (2009 – 2018) airline delay and cancellation data will be made to determine the causes of airline delays.

# A3. Summary of Published Works

Various factors contribute to aircraft delays. One of these factors is the impact that weather, especially extreme weather, can have in delaying flights. In January 2019, an article titled, “Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States,” written by Christopher J. Goodman and Jennifer D. Small Griswold detailed the average impact that certain weather phenomena have on aircraft delays. They found that it is important to assess airports individually as their operations efficiencies are impacted due to differences in weather and airline operations. Their findings included:

1. Differences in airport weather and airline operations impact the efficiencies of airport operations and it is important to assess airports individually.
2. Understanding the differences in airport weather climates allows for an understanding of how inclement weather reduces efficiency.
3. Weather impacts delays and cancellations in a way consistent with climatological weather patterns.

Nicholas G. Rupp writes, in article titled, “Further Investigations into the Causes of Flight Delays,” that flight delays should be investigated from both the airline and passenger perspectives. Three conclusions are reported:

1. There are considerable differences when using alternative measures of flight delays – excess travel time which is the airline perspective and minutes of arrival/departure delay which is the passenger perspective.
2. Seating capacity, load factor, departure time, and distance have significant effects on flight delays.
3. After controlling for airport-specific effects, most estimations indicate that airport concentration at origination has longer departure and arrival delays.

Paul Blackwood wrote, in an article titled, “Understanding Flight Delays at U.S. Airports in 2010, Using Chicago O’Hare International Airport as a Case Study,” that flight delays negatively impact the environment, the economy, and society. After studying various delays at the named airport, he concluded:

1. Late aircraft accounted for 40% of all delays.
2. Carrier related issues accounted for 29% of all delays.
3. National airspace bottlenecks accounted for 25% of all delays.
4. Weather accounted for 6% of all delays.
5. Security issues accounted for 1% of all delays.

# A3a. Relation of Published Works to Project

It is possible that weather one of the causes of flight delays, or it could be the largest cause of flight delays. This project must look at delays caused by the weather and determine how much effect weather has on delays.

Delays will be investigated more from the passenger side of things in this analysis, as this is the biggest of people who are directly affected by delays.

It will be interesting to see if the causes of delays at a single large airport are indicative of delays across the entire country.

# A4. Summary of Data Analytics Solution

This project will analyze airline delay and cancellation data from 2009 to 2018 and determine the causes of flight delays. Flight delays cost the US economy a great deal of money ($32.9 billion in 2007) and determining the cause of these delays could allow this cost to be mitigated or reduced.

# A5. Benefit to Organization and Decision-Making Process

The benefits of this analysis will help both airlines and passengers. For airlines, the causes of delays will be identified, and possible mitigation strategies can be formulated. For passengers, the causes of delays may help them to determine better times to fly or which airlines offer the least delays.

# Data Analytics Plan

# B1. Goals, Objectives, and Deliverables

The goal of this project is to create a Jupyter Notebook to perform data analysis on flight delay and cancellation data from 2009 – 2018 in order to discover the cause(s) of flight delays.

The objectives for this goal are:

* Concatenate the data into a single dataset, so that data analysis can be performed on a single dataset.
  + The deliverable is to return a single dataset containing all the years of data.
* Cleanse the dataset, so that missing or unknown data does not compromise the results.
  + The deliverable is to return a single dataset free from unknown or missing data.
* Analyze the dataset for the cause(s) of flight delays.
  + The deliverable is to find the causes of flight delays.

# B2. Scope of Project

The scope of this project will include a Jupyter notebook. This notebook will provide various steps in order to analyze the flight delay and cancellation dataset. The output of this notebook will be the results of each step culminating in the cause(s) of flight delays. The scope of this project will not include analyzing cancelled or diverted flights.

# B3. Standard Methodology

This project will utilize a Waterfall project methodology. This consists of five steps where the first must be completed before the next can be undertaken. Those five steps are Requirements, Design, Implementation, Verification, and Maintenance. The following explains how I plan to proceed during each of these phases.

**Requirements:** One of the key aspects of the Waterfall methodology is that all customer requirements are gathered initially. In this step, I will determine the project scope, the user expectations, and the resources needed to complete the project.

**Design:** In this stage of the methodology, I will compile the tasks needed to be completed to achieve the project objectives. Some of these tasks include determining what data cleansing will be necessary, the steps needed to analyze the dataset, and the visualizations for the results.

**Implementation:** In this stage, I will complete the tasks needed to achieve the objectives and test the Jupyter notebook to ensure it is producing the desired results.

**Verification:** In this stage, I will complete a standalone file for this project, so that it can be implemented by anyone else who has access to suitable hardware and software, i.e., Conda installed.

**Maintenance:** This stage will not apply to this project, as it will not be in production in any companies. However, it could be uploaded to sites such as [www.kaggle.com](http://www.kaggle.com), and in that case bug fixes and modifications could be requested.

# B4. Timeline and Milestones

Present a table showing for each milestone its projected start and end dates, and its projected duration:

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected End Date** | **Duration (hours)** |
| Establish requirements for analytics process | 03/01/2024 | 03/03/2024 | 24 |
| Download dataset | 03/04/2024 | 03/04/2024 | 2 |
| Code notebook – loading data | 03/04/2024 | 03/04/2024 | 6 |
| Code notebook – cleansing data | 03/05/2024 | 03/08/2024 | 32 |
| Code notebook – data analysis | 03/09/2024 | 03/14/2024 | 48 |
| Test notebook | 03/15/2024 | 03/18/2024 | 32 |
| Create html file showing all notebook code and results | 03/19/2024 | 03/19/2024 | 1 |

# B5. Resources and Costs

|  |  |
| --- | --- |
| **Personnel, technology, or infrastructure** | **Cost** |
| Dataset for data analysis | N/A |
| Jupyter notebook development environment | N/A |
| 145 hours | N/A |

The resources needed for this project to be completed and implemented are limited to a dataset for analysis (downloaded for free from www.kaggle.com), my time (120 hours of conception, design, coding, and testing), my computer for development (no cost, it has already been acquired). There are no additional resources or costs that will be associated with this project.

# B6. Criteria for Success

How will you measure the success and effectiveness of outcomes? Present at least 3 criteria for success, including how you will collect the data for each criterion and what constitutes success. Summarize your discussion in a table:

|  |  |  |
| --- | --- | --- |
| **Criterion/Metric** | **Required Data** | **Cut Score for Success** |
| Were reasons for flight delays identified? | Output of Jupyter notebook | Success is if and only if there is a correlation for each reason |
| Identification of main cause of flight delays. | Output of Jupyter notebook | Success is if and only if there is a main cause of flight delays |
| Identify main cause of flight delays for airline with greatest delay. | Output of Jupyter notebook | Success is if and only if an airline with greatest delay has a main cause of flight delay |
| Identify main cause of flight delays for day of the week with greatest delays. | Output of Jupyter notebook | Success if and only if a day of the week is identified that has greatest delay |

# Design of Data Analytics Solution

# C1. Hypothesis

Determine the main cause of flight delays.

# C2. Analytical Method

The analytical method I will use in my data analysis will be descriptive. The data analysis technique I will use is multivariate analysis in the form of a correlation matrix.

# C2a. Justification of Analytical Method

I chose this method and technique because there are several possible variables that may depend on each other. The correlation matrix enables those variables that are correlated to be identified easily. This enables the causes of flight delays to be identified and ranked.

# C3. Tools and Environments of Solution

A Jupyter notebook will be used to manipulate the data, because it provides a robust environment to manipulate and analyze the dataset. This tool also provides a flexible way to output results as both text and graphical plots can be generated.

# C4. Methods and Metrics to Evaluate Statistical Significance

Statistical significance for the correlation coefficients of flight delay causes will be shown via a t-test. A null hypothesis will be proposed for each flight delay cause, and an attempt will be made to disprove that hypothesis, thereby proving that correlation coefficient is statistically significant. P-value of less than 0.05 will show that the result is significant.

# C4a. Justification of Methods and Metrics

This approach is the most appropriate because it will allow me to demonstrate whether a flight delay cause is statistically significant.

# C5. Practical Significance

The practical significance of knowing what causes flight delays allows mitigation strategies to be formulated, and whether attempting mitigation is useful. For example, if weather was the biggest cause of flight delays, then it would be impractical to try and mitigate the effect of weather on delays. It would then be more productive to try and mitigate other causes of delays. Delays that are under an airline’s control could be reduced to provide better customer service and goodwill. For example, allowing planes to fly faster when late would help reduce or negate late arrivals.

# C6. Visual Communication

A correlation matrix heatmap will be used to communicate the findings, as well as a bar plot of the number of delays versus the date of each flight. The heatmap is very good at displaying many correlation coefficients compared to other methods in a fixed space. The bar plot is a very good method of comparing the number of delays for each delay type in a fixed space.

# Description of Datasets

# D1. Source of Data

Airline Delay and Cancellation Data data were collected from datasets provided by [www.kaggle.com](http://www.kaggle.com). This website provides various publicly downloadable datasets. This data has been combined from multiple US Government (Bureau of Transportation Statistics) datasets.

There are ten data files associated with this project:

2009.csv: Airline Delay and Cancellation Data dataset for 2009 (Milan, 2020).

2010.csv: Airline Delay and Cancellation Data dataset for 2010 (Milan, 2020).

2011.csv: Airline Delay and Cancellation Data dataset for 2011 (Milan, 2020).

2012.csv: Airline Delay and Cancellation Data dataset for 2012 (Milan, 2020).

2013.csv: Airline Delay and Cancellation Data dataset for 2013 (Milan, 2020).

2014.csv: Airline Delay and Cancellation Data dataset for 2014 (Milan, 2020).

2015.csv: Airline Delay and Cancellation Data dataset for 2015 (Milan, 2020).

2016.csv: Airline Delay and Cancellation Data dataset for 2016 (Milan, 2020).

2017.csv: Airline Delay and Cancellation Data dataset for 2017 (Milan, 2020).

2018.csv: Airline Delay and Cancellation Data dataset for 2018 (Milan, 2020).

# D2. Appropriateness of Dataset

These datasets are appropriate for addressing the project goal since they contain the proper fields for analysis. Using these datasets is the most appropriate way to provide my analysis with the data as they are freely provided by the US Government. This means that I will not need to use any propriety data and any legal ramifications that entails.

# D3. Data Collection Methods

The data was collected by downloading it from Kaggle. The advantages of collecting data this way is that some data cleansing had taken place, since this data had ultimately come from US Government sources, and it had some modifications made. The disadvantages of collecting data this way is that some of the data was not needed, so further cleansing was required. The data was examined and cleansed to ensure that it was as accurate as possible and of high quality, i.e., no missing data was used to contaminate the results. An account with Kaggle had to be created before the data could be downloaded.

# D4. Data Quality

The data had some quality issues. One column of data was unusable. It had no useful column name and contained no actual data. This column was removed from the dataset. Some data that required a zero value did not contain that value and that had to be corrected. Some data was assigned the wrong type when loaded. This was corrected.

# D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

The data that were used are publicly available to anyone, and do not contain any personally identifiable information. While the data should be under the control of internal standards for data governance, there are few issues that would be caused if the data was stolen, since it is already publicly available.

# D5a. Precautions

When the dataset was modified, tests were performed to confirm that the modification was correctly executed. During analysis, each step was checked to make sure that the results obtained for each step were correct. The project files (data files, Jupyter notebook, PDF files) were stored on local storage, but at the end of each day were pushed to GitHub in case of a local issue. The project files are locally stored under a user account that is password protected.

# Sources

Rupp, N. G. (2007, May 20). *Further Investigations into the Causes of Flight Delays*. East Carolina University. <https://economics.ecu.edu/wp-content/pv-uploads/sites/165/2019/07/ecu0707.pdf>

Blackwood, P. (2012, April). *Understanding Flight Delays at U.S. Airports in 2010, Using Chicago O’Hare International Airport as a Case Study*. Western Michigan University <https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1033&context=masters_theses>

Goodman, C. J. and Small Griswold, J. D. (2019, January 4). *Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States*. Journal of Applied Meteorology and Climatology. <https://journals.ametsoc.org/view/journals/apme/58/3/jamc-d-17-0277.1.xml?tab_body=pdf>

Milan, T. (2020). *Airline Delay and Cancellation Data 2018.* Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2009.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2010.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2011.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2012.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*.Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2013.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2014.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2015.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2016.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2017.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2018.csv